Unsupervised Identification of Translationese

Ella Rabinovich
Department of Computer Science
University of Haifa
ellarabi@csweb.haifa.ac.il

Shuly Wintner
Department of Computer Science
University of Haifa
shuly@cs.haifa.ac.il

Abstract

Translated texts are distinctively different from original ones, to the extent that supervised text classification methods can distinguish between them with high accuracy. These differences were proven useful for statistical machine translation. However, it has been suggested that the accuracy of translation detection deteriorates when the classifier is evaluated outside the domain it was trained on. We show that this is indeed the case, in a variety of evaluation scenarios. We then show that *unsupervised* classification is highly accurate on this task. We suggest a method for determining the correct labels of the clustering outcomes, and then use the labels for voting, improving the accuracy even further. Moreover, we suggest a simple method for clustering in the challenging case of mixed-domain datasets, in spite of the dominance of domain-related features over translation-related ones. The result is an effective, fully-unsupervised method for distinguishing between original and translated texts that can be applied to new domains with reasonable accuracy.

1 Introduction

Human-translated texts (in any language) have distinct features that distinguish them from original, non-translated texts. These differences stem either from the effect of the translation process on the translated outcomes, or from "fingerprints" of the source language on the target language product. The term *translationese* was coined to indicate the unique properties of translations.

Awareness to translationese can improve statistical machine translation (SMT). First, for training translation models, parallel texts that were translated in the direction of the SMT task are preferable to texts translated in the opposite direction; second, for training language models, monolingual corpora of translated texts are better than original texts.

It is possible to automatically distinguish between original (O) and translated (T) texts, with very high accuracy, by employing text classification methods. Existing approaches, however, only employ *supervised* machine-learning; they therefore suffer from two main drawbacks: (i) they inherently depend on data annotated with the translation direction, and (ii) they may not be generalized to unseen (related or unrelated) domains.[1] These shortcomings undermine the usability of supervised methods for translationese identification in a typical real-life scenario, where no labelled in-domain data are available.

In this work we explore *unsupervised* techniques for reliable discrimination of original and translated texts. More precisely, we apply *dimension reduction* and *centroid-based clustering* methods (enhanced by internal clustering evaluation), for telling O from T in an unsupervised scenario. Furthermore, we introduce a robust methodology for labelling the obtained clusters, i.e., annotating them as "original" or "translated", by inspecting similarities between the clustering outcomes and O and T *prototypical* examples. Rigorous experiments with four diverse corpora demonstrate that clustering of in-domain texts using lexical, content-independent features systematically yields very high accuracy, only 10 percent points lower than the performance of supervised classification on the same data (in most cases). Ac-

[1] We use "domain" rather freely henceforth to indicate not only the topic of a corpus but also its modality (written vs. spoken), register, genre, date, etc.

Transactions of the Association for Computational Linguistics, vol. 3, pp. 419–432, 2015. Action Editor: Philipp Koehn.
Submission batch: 4/2015; Revision batch 7/2015; Published 7/2015.

curacy can be improved even further by *clustering consensus* techniques.

We further scrutinize the tension between domain-related and translationese-based text properties. Using a series of experiments in a *mixed-domain* setup, we show that clustering (in particular, relying on content-independent features) perfectly groups the data into domains, rather than into the (desirable) cross-domain O and T; that is, domain-related properties clearly dominate and overshadow the translationese-based characteristics of the underlying texts. We address the challenge of discriminating O from T in a mixed-domain setup by proposing two simple methodologies (*flat* and *two-phase*) and empirically demonstrate their soundness.

The clustering experiments throughout the paper were conducted in a setup similar to that of supervised classification, determining the status (O vs. T) of logical units (chunks) of 2,000 tokens. We also show that clustering accuracy remains stable even when the number of available chunks decreases dramatically and remains satisfactory when the chunk size is reduced.

The main contribution of this work is therefore two-fold: (i) we establish a robust approach for reliable unsupervised identification of translated texts, thereby eliminating the need for in-domain labeled data; (ii) we provide an extensive empirical foundation for the dominance of domain-based properties over translationese-related characteristics of a text, and propose a methodology for identification of translationese in a mixed-domain scenario.

The remainder of the paper is structured as following: after reviewing related work in Section 2, we detail our datasets, features and tools in Section 3. In Section 4 we reproduce and extend supervised classification results, and demonstrate the poor cross-domain classification accuracy of supervised methods. Our clustering methodology and experiments are described in Section 5; mixed-domain classification is discussed in Section 6. We conclude with a discussion and suggestions for future research.

2 Related Work

Much research in Translation Studies indicates that translated texts have unique characteristics. Translated texts (in any language) constitute a sub-language (sometimes referred to as a *genre*, or a *dialect*) of the target language, presumably reflecting both the artifacts of the translation process and traces of the original language from which the texts were translated (the *source* language). Gellerstam (1986) called this sub-language *translationese*, and suggested that the differences between O and T do not indicate poor translation but rather a *statistical phenomenon*, caused by a systematic influence of the source language on the target language.

These differences have ramifications for SMT. Kurokawa et al. (2009) were the first to note it: they showed that *translation models* trained on English-translated-to-French bitexts were much better than ones trained on French-translated-to-English, when the SMT task is translating English to French. Lembersky et al. (2012a, 2013) corroborated these results, for more language pairs, and suggested a way to adapt translation models to the properties of translationese. Furthermore, Lembersky et al. (2011, 2012b) showed that *language models* compiled from translated texts are better for SMT than ones compiled from original texts. These results all highlight the practical importance of being able to reliably distinguish between translated and original texts.

Indeed, translated texts are so markedly different from original ones that automatic classification can identify them with very high accuracy (Baroni and Bernardini, 2006; Ilisei et al., 2010; Ilisei and Inkpen, 2011; Popescu, 2011). Recently, Volansky et al. (2015) investigated several translation studies hypotheses by performing an extensive exploration of the ability of various feature sets to distinguish between O and T. Using SVM classifiers and ten-fold cross-validation evaluation, they listed several features that yield near perfect accuracy.

Most works mentioned above train and evaluate classifiers on texts drawn from the same corpus. When these classifiers are tested on texts from different domains, or in a different genre, or translated from a different language, classification accuracy dramatically deteriorates. Koppel and Ordan (2011) train classifiers on the Europarl corpus (Koehn, 2005), with English translated from five different languages. When the classifiers are evaluated on English translated from the same language they were trained on, accuracy is near 100%; but when evaluated on translations from a different language, accuracy drops significantly, in some cases below 60%. This pattern recurs when the test corpus is different from the training corpus (newspaper articles vs. parliament proceedings). Similarly, Avner et al. (Forthcoming) report excellent (near 100%) results identifying Hebrew translationese on a corpus of literary texts, using very simple word-level features. Evaluation on different domains (popular science) and on Hebrew translated from French, rather than English, however, shows much poorer results, with accuracies around 60% in many cases.

We hypothesize that the main reason for the deterioration in the accuracy of (supervised) translationese classifiers when evaluated out-of-domain stems from the fact that domain differences overshadow the differences between O and T. Diwersy et al. (2014) studied various sorts of linguistic variation by applying semi-supervised multivariate techniques. They investigated, among other factors, register variation in English and German originals and translations. By applying a series of supervised and unsupervised statistical analyses, they demonstrated that register-related properties are much better exhibited by the underlying texts than properties related to the documents' translation status. We address the challenge of mixed-domain classification in Section 6.

One way to overcome the dependence on labeled data and domain-overfitting of supervised classifiers is to use *unsupervised* methods, in particular *clustering*. The only application of clustering to translationese that we are aware of is the work of Nisioi and Dinu (2013), who investigated translationese- and authorship-related characteristics by applying hierarchical clustering to books written by a Russian-English bilingual author. While they mainly focused on authorship attribution, Nisioi and Dinu (2013) also demonstrated that it is possible to discriminate O from T by applying clustering with lexical features (function words) extracted from complete books (25,000–180,000 tokens). We address the challenge of unsupervised identification of translationese using a different methodology and much smaller logical units (2,000 tokens), and further broaden the scope of our work by proposing a methodology for telling O from T in mixed-domain scenarios.

Unsupervised classification is a well-established discipline; in this work we use *KMeans* (Lloyd, 1982) for clustering and *KMeans++* (Arthur and Vassilvitskii, 2007) as a KMeans initialization method.

3 Experimental Setup

3.1 Datasets

Our main dataset[2] consists of texts originally written in English and texts translated to English from French. We use various corpora: (i) Europarl, the proceedings of the European Parliament (Koehn, 2005), between the years 2001-2006; (ii) the Canadian Hansard, transcripts of the Canadian Parliament, spanning years 2001-2009; (iii) literary classics written (or translated) mainly in the 19th century; and (iv) transcripts of TED and TEDx talks. This collection suggests diversity in genre, register, modality (written vs. spoken) and era. Table 1 details some statistical data on the corpora (after tokenization).[3] We now briefly describe each dataset.

Europarl is probably the most popular parallel corpus in natural language processing, and it was indeed used for many of the translationese tasks surveyed in Section 2. This corpus has been used extensively in SMT (Koehn et al., 2009), and was even adapted specifically for research in translation studies: Islam and Mehler (2012) compiled a customized version of Europarl, where the direction of translation is indicated. We use a version of Europarl (Rabinovich and Wintner, Forthcoming) that aims to further increase the confidence in the direction of translation, through a comprehensive cross-lingual validation of the original language of the speakers.

The Hansard is a parallel corpus consisting of transcriptions of the Canadian parliament in English and French between 2001 and 2009. This is the largest available source of English–French sentence pairs. We use a version that is annotated with the original language of each parallel sentence. Relying on metadata available in the corpus, we filtered out all segments not referring to speech, i.e., retaining only sentences annotated as *Content ParaText*.

[2] The dataset is available at `http://cl.haifa.ac.il/projects/translationese`.

[3] We use "EUR", "HAN", "LIT" and "TED" to denote the four corpora in the discussion below.

	Number of sentences			Number of tokens		Number of types	
Corpus	Original E	F→E	Total	Original E	F→E	Original E	F→E
EUR	134,725	71,816	206,541	3,406,513	2,112,085	37,203	28,119
HAN	3,441,984	757,573	4,199,557	65,491,960	13,457,613	158,645	63,192
LIT	36,123	85,210	121,333	858,297	1,750,525	25,113	38,842
TED	7,551	4,827	12,378	129,334	87,214	9,667	7,441

Table 1: Corpus statistics

The Literature corpus consists of literary classics written (and translated) in the 18th–20th centuries by English and French authors; the raw material is available from the Gutenberg project. We use subsets that were manually or automatically paragraph-aligned. Note that classifying literary texts is considered a more challenging task than classifying more "technical" translations, such as parliament proceedings, since translators of literature typically enjoy more literary freedom, thereby rendering the translation product more similar to original writing (Lynch and Vogel, 2012; Avner et al., Forthcoming).

Our TED talks corpus consists of talks originally given in English and talks translated to English from French. The quality of translations in this corpus is very high: not only are translators assumed to be competent, but the common practice is that each translation passes through a review before being published. This corpus consists of talks delivered orally, but we assume that they were meticulously prepared, so the language is not spontaneous but rather planned. Compared to the other sub-corpora, the TED dataset has some unique characteristics that stem from the following reasons: (i) its size is relatively small; (ii) it exhibits stylistic disparity between the original and translated texts (the former contains more "oral" markers of a spoken language, while the latter is a written translation); and finally (iii) TED talks are not transcribed but are rather subtitled, so they undergo some editing and rephrasing.[4]

The vast majority of TED talks are publicly available online, which makes this corpus easily extendable for future research.

3.2 Processing and Tools

All datasets are first tokenized using the Stanford tools (Manning et al., 2014) and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary). We assume that translationese-related features are present in the texts across author or speaker, thus we allow some chunks to contain linguistic information from two or more different texts simultaneously. For the main (single-corpus) classification experiments we use 2000 text chunks each from Europarl and Hansard, 800 from Literature and 88 chunks from TED; each sub-corpus consists of an equal number of original and translated chunks. For every classification experiment we use the maximal equal number of chunks from each class, thus we always (randomly) down-sample the datasets in order to have a comparable number of training/testing examples for supervised classification, and comparable cluster size for clustering.

We use Weka (Hall et al., 2009) as the main tool for classification, clustering, and dimension reduction. In all the classification experiments, we use SVM (SMO) as the classification algorithm with the default linear kernel. For clustering we use Weka's KMeans implementation (SimpleKMeans) with the KMeans++ initialization strategy. We use Eucledian distance as the similarity measure for KMeans, and apply a custom clustering-evaluation-based wrapper (see Section 5) to further enhance Weka's basic clustering implementation.

We use Principal Component Analysis (PCA, Jolliffe (2002)) for dimension reduction. PCA is a statistical procedure that discovers variables with the largest possible variance, i.e., features that account for most variability in the data (*principal components*). It performs a linear mapping of the data to a lower-dimensional space in a way that maximizes the variance of the data in the low-dimensional representation, by removing highly correlated or superfluous variables. The outcome of PCA is a new set of features, each of which is a linear combination of the discovered components. The number of the newly

[4] http://translations.ted.org/wiki/How_to_Compress_Subtitles

generated variables varies from one to the number of variables originally used to describe the data, and is typically controlled by a parameter.

Apart from the enhanced efficiency (due to the reduced computational costs), dimensionality reduction often carries a positive effect on the accuracy of the underlying classification task, especially when the data are meager or feature vectors are sparse. The (accuracy-wise) optimization gains of PCA, when followed by the KMeans clustering algorithm, were reported by Ng et al. (2001). We perform dimension reduction using the Weka implementation of PCA, with the "variance_covered" parameter set to 0.1 across all feature types and datasets, prior to applying a clustering procedure.

3.3 Features

We focus on a set of features that reflect lexical and structural properties of the text, and have been shown to be effective for supervised classification of translationese (Volansky et al., 2015). Specifically, we use *function words* (FW), more precisely, the same list that was used in previous works on classification of translationese (Koppel and Ordan, 2011; Volansky et al., 2015). Feature values are raw counts (further denoted by *term frequency, tf*), normalized by the number of tokens in the chunk; the chunk size may slightly vary, since the chunks respect sentence boundaries. For the clustering experiments we further scale the normalized *tf* by the *inverse document frequency (idf)*, which offsets the importance of a term by a factor proportional to its frequency in the corpus. The *tf-idf* statistic has been shown to be effective with *lexical* features, and is often used as a weighting factor in information retrieval and text mining. While function words are assumed to be very frequent, their counts within a text vary greatly (e.g., "the" vs. "whereas"). We therefore opt for *tf-idf* weighting of FW across all sub-corpora.

In addition to function words, we experiment with several other feature sets, including character trigrams, part-of-speech (POS) trigrams, *contextual function words* and *cohesive markers*. Contextual function words are a variation of POS trigrams where a trigram can be anchored by specific function words: these are consecutive triplets $\langle w_1, w_2, w_3 \rangle$ where at least two of the elements are function words, and at most one is a POS tag. Cohesive markers are words or phrases that signal the underlying flow of thought: they organize a composition of phrases by specifying the type, purpose or direction of upcoming ideas, and can therefore serve as evidence of the translation process. We use the list of 40 cohesive markers defined in Volansky et al. (2015).

Character, POS, and contextual FW trigrams are calculated as detailed in Volansky et al. (2015), but we only consider the 1000 most frequent feature values extracted from each dataset (or a combination of datasets) being classified. This subset yields the same classification quality as the full set, reducing computation complexity.

4 Supervised Classification

We begin with supervised classification, reestablishing the high accuracy of in-domain (supervised) classification of translationese, but highlighting the deterioration in accuracy when cross-domain classification is considered. We first reproduce the Europarl classification results with the best performing feature sets, as reported by Volansky et al. (2015), and present results for three additional sub-corpora: Hansard, Literature and TED. Table 2 lists the ten-fold cross-validation classification accuracy with various features. All features (except perhaps cohesive markers) yield excellent accuracy.

feature / corpus	**EUR**	**HAN**	**LIT**	**TED**
FW	96.3	98.1	97.3	97.7
char-trigrams	98.8	97.1	99.5	100.0
POS-trigrams	98.5	97.2	98.7	92.0
contextual FW	95.2	96.8	94.1	86.3
cohesive markers	83.6	86.9	78.6	81.8

Table 2: In-domain (cross-validation) classification accuracy using various feature sets

A few previous works suggested that cross-domain classification of translationese results in low accuracy (Koppel and Ordan, 2011; Avner et al., Forthcoming). Our experiments corroborate this observation; Table 3 depicts the cross-domain classification accuracy on the Europarl, Hansard and Literature corpora, when training on one corpus and testing on another (using function words).[5] A balanced setup for this experiment was generated by randomly selecting 800 chunks from each corpus, divided equally to O and T. The results only slightly outperform chance level, even for the Europarl–Hansard seemingly domain-related pair: we obtain 59.7% to 60.8% accuracy in the two directions.

train / test	**EUR**	**HAN**	**LIT**	**10-fold x-validation**
EUR		60.8	56.2	94.7
HAN	59.7		58.7	98.1
LIT	64.3	61.5		97.3

Table 3: Pairwise cross-domain classification using function words

Attempting to enrich the classifier's training "experience" we conducted additional experiments, where we train on two sub-corpora out of Europarl, Hansard and Literature, and test on the remaining one. The results are depicted in Table 4. Here, too, accuracy is very low, implying that training on diverse data does not necessarily provide a solution for cross-domain classification of translationese. The right-hand column of the table reports ten-fold cross-validation results of the two sub-corpora that are subject for training. Excellent in-domain classification results on the one hand and poor cross-domain predictive performance on the other, imply that the model describing the relation in a certain domain is inapplicable to a different (even seemingly similar) domain due to significant differences in the distribution of the underlying data.

train / test	**EUR**	**HAN**	**LIT**	**10-fold x-validation**
EUR + HAN			63.8	94.0
EUR + LIT		64.1		92.9
HAN + LIT	59.8			96.0

Table 4: Leave-one-out cross-domain classification using function words

Reflecting the poor generalization capability of translationese features, these results call for developing other methodologies for reliably discriminating O from T, specifically, methodologies that are independent of in-domain labeled data.

5 Clustering

5.1 Initial results

To overcome the domain-dependence of supervised classification, we experiment in this section with unsupervised methods. We begin with the KMeans clustering algorithm, using KMeans++ initialization policy and dimension reduction. To evaluate the accuracy of the algorithms, each cluster is labeled by the majority of (O or T) instances it includes (using ground truth annotations), and the overall precision is the percentage of instances correctly assigned to their respective clusters (we discuss *unsupervised* cluster labeling in Section 5.2).

The KMeans clustering algorithm (with any initialization policy) is sensitive to the initial settings of its parameters, in particular the initial choice of centroids. A cluster *centroid* is the geometrical center of all observations within the cluster. The result of the KMeans algorithm may significantly vary according to its first step: the initial assignment of (random) points to cluster centroids. We address this potential pitfall by performing N clustering iterations, randomly varying the initial parameter settings, outputting the outcome that exhibits the highest similarity of points within a cluster. Formally, let C_i^j denote cluster i in iteration j, and let m_i^j denote this cluster's centroid, so that $i\in$[1,2], and $j\in$[1..N]. *Sum-of-Square-Error (SSE)* is an intrinsic clustering evaluation metric that measures the similarity of elements in a cluster. The SSE of C_i^j is defined by

$$SSE_i^j = \sum_{x \in C_i^j} (x - m_i^j)^2$$

We aim to optimize the clustering result by choosing an outcome that minimizes the accumulative SSE:

$$\arg\min_{j} SSE^j = \arg\min_{j\in[1..N]} \sum_{i\in[1,2]} SSE_i^j$$

The selected clustering outcome represents the result of a *single* clustering experiment. The described method for selecting a clustering outcome can be viewed as a binary version of the *Bisecting* KMeans

[5] We focus mainly on function words, because they are known to reflect stylistic differences rather than contents or specific corpus features, and are therefore less susceptible to domain overfitting. Other feature sets yielded similar results.

algorithm; it is applied in all experiments throughout the paper, with number of iterations (N) fixed to 5, following the recommendation by Steinbach et al. (2000, p. 13).

We conducted a series of experiments with various feature sets; the main results are depicted in Table 5. The reported numbers reflect the average accuracy over 30 experiments (the only difference being a random choice of the initial conditions).[6]

feature / corpus	**EUR**	**HAN**	**LIT**	**TED**
FW	88.6	88.9	**78.8**	**87.5**
char-trigrams	72.1	63.8	70.3	78.6
POS-trigrams	**96.9**	76.0	70.7	76.1
contextual FW	92.9	**93.2**	68.2	67.0
cohesive markers	63.1	81.2	67.1	63.0

Table 5: Clustering results using various feature sets

First and foremost, the results are very good, ranging from a few percent points lower than supervised classification (Table 2, Europarl and Hansard) to approximately 25 percent points lower in a few cases (e.g., Literature). Function words systematically yield very high accuracy; the quality of clustering with other features varies across the sub-corpora. Cohesive markers perform poorly (with a single exception, Hansard), which mirrors the moderate supervised classification precision achieved with the same feature set.

The exceptionally high result of Europarl with POS-trigrams can be attributed to the excessive frequency of specific phrases in the translated Europarl texts (in contrast to their original counterparts).[7] We explain the lower precision achieved on the Literature corpus by its diverse character: it comprises works attributed to a variety of authors, periods and genres, which is challenging for the unsupervised algorithm (see Section 6). A notably high accuracy is obtained on the small TED corpus, which implies the applicability of our clustering methodology to data-meager scenarios.

We conducted an additional set of experiments with unequal proportions of original and translated texts, considering twice the number of O chunks compared to T and vice versa. The average clustering accuracy using FW is similar to that obtained in the balanced setup (Table 5): 87.5% on Europarl, 88.9% on Hansard, 73.2% on Literature, and 88.6% on the TED sub-corpus.

5.2 Cluster labeling

As is always the case with unsupervised methods, clustering can divide observations into classes but cannot label those classes. A *cluster labeling* algorithm examines the contents of each cluster in order to find labels that best summarize its members, and distinguish the clusters from each other.

In the context of translationese identification, the task of cluster labeling is to determine which of the produced clusters represents O, and which T. We address this challenge by exploring similarities between the *language models* of the obtained clusters, and language models of (presumably) *prototypical* O and T samples. A simple unigram language model assigns each word a probability proportional to its frequency in the underlying text; we use smoothed term frequencies scaled by the inverse total term frequencies. We then compare language models to reveal similarities between the prototypical O and T samples and the chunk sets produced by clustering.

The construction method of prototypical LMs is motivated by (i) abstracting from content, by utilizing only function words for this purpose; and (ii) attempting to avoid the interference of domain-related properties, by considering only (presumably) *universal* markers: words that share similar frequency patterns in several datasets w.r.t. to O vs. T.

Let O_m (O-markers) denote a set of function words that tend to be associated with O. We select this set by picking words whose frequency in O is excessive, compared to T; more precisely, the ratio of their frequency in O and T is above (1+δ), where δ=0.05. Similarly, T_m (T-markers) is a set of words with O-to-T frequency ratio below (1-δ). We create a prototypical O example by the concatenation of O_m, and a prototypical T example by the concatenation of T_m. The language model of these examples is then constructed by the ϵ-smoothed likelihood of each term in the markers vocabulary $V = O_m \bigcup T_m$, where ϵ=0.001.

[6] Standard deviation in most experiments was close to 0.

[7] As an example (and in line with van Halteren (2008)), in the 2000 Europarl chunks, the phrase *ladies and gentlemen* appears 1258 times in T, but only 12 times in O.

Formally, for $w \in V$,

$$p(w \mid O_m) = \frac{tf(w) + \epsilon}{|O_m| + \epsilon \times |V|}$$
$$p(w \mid T_m) = \frac{tf(w) + \epsilon}{|T_m| + \epsilon \times |V|}$$

We denote the resulting language models by P_O and P_T, respectively. Given two clusters, C_1 and C_2, we similarly compute their language models, denoted by P_{C_1} and P_{C_2}, respectively, over the vocabulary V. We measure the similarity between a class X (either O or T) and a cluster C_i using the Jensen-Shannon divergence (JSD) (Lin, 1991) on the respective probability distributions. Specifically, we define the *distance* between the language models as the square root of the divergence value, which is a metric, often referred to as *Jensen-Shannon distance* (Endres and Schindelin, 2003):

$$D_{JS}(X, C_i) = \sqrt[2]{JSD(P_X || P_{C_i})}$$

The assignment of the label X to the cluster C_1 is then supported by both C_1's proximity to the class X and C_2's proximity to the other class:

$$label(C_1) = \begin{cases} \text{“O”} & \text{if } D_{JS}(O, C_1) \times D_{JS}(T, C_2) < \\ & \alpha \times D_{JS}(O, C_2) \times D_{JS}(T, C_1) \\ \text{“T”} & \text{otherwise} \end{cases}$$

C_2 is assigned the complementary label. The value of α is fixed to 1 in this equation, but we note that it can be varied for further investigation of the relatedness of the underlying language models.

We apply the cluster labeling technique described above to determine the labels of generated clusters. We construct prototypical O- and T-texts by selecting O- and T-markers from a random sample of Europarl and Hansard texts, using 600 chunks from each corpus.[8] We then compare the language models induced by these samples to those of the generated clusters (tested on different chunks, of course) to determine the cluster labels; the predicted labels are then verified against the majority-driven labeling, based on ground truth annotations. We apply this procedure to the outcome of all clustering experiments (per domain, using various features), achieving overall precision of 100%. In other words, the labeling procedure yields prefect accuracy not only on Europarl and Hansard texts that were not used for generation of O and T prototypical examples, but also on unseen Literature and TED datasets. We conclude that it is possible, in general, to determine the labels of clusters produced by our clustering algorithm with perfect accuracy.

5.3 Clustering consensus among feature sets

Since different feature sets have different predictions on our data, we hypothesize that consensus voting can improve the accuracy of clustering. We treat each individual clustering result (based on a certain feature set) as a judge, voting whether a single text chunk belongs to O or to T. We use the cluster labeling method of Section 5.2 to determine labels. The final assignment of a label to a cluster is determined by the majority vote of the various judges.

Table 6 presents the results of these experiments. We compare consensus results to the accuracy achieved by function words, the best-performing single feature set (on average), see Table 5. Both three judges and five judges yield a consistent increase in accuracy. Five judges systematically (and, on Europarl and Hansard, significantly) outperform the result of clustering with functions words only. This indicates that various features tend to capture different aspects of translationese, that are eventually leveraged by the “fusion” of different clustering results into a single, higher-quality outcome.

5.4 Sensitivity analysis

In supervised classification, the amount of labeled data has a critical effect on the classification accuracy. This does not seem to be the case with clustering: accuracy remains stable when the number of chunks used for classification decreases (Figure 1a). Evidently, as few as 300 chunks are sufficient for excellent classification.[9] We attribute the (slight) fluctuations in the graph to the random choice of the subset of chunks that are subject for clustering. Naturally, clustering accuracy stabilizes when the number of chunks increases, since the effect of random

[8] This subset of the Europarl and Hansard corpora was used for one-time generation of prototypical O and T language models, and excluded from further use.

[9] The results on the Literature corpus are limited by the amount of available data in this dataset.

method / corpus	**EUR**	**HAN**	**LIT**	**TED**
FW	88.6	88.9	78.8	87.5
FW char-trigrams POS-trigrams	91.1*	86.2	78.2	**90.9***
FW POS-trigrams contextual FW	**95.8***	89.8	72.3	86.3
FW char-trigrams POS-trigrams contextual FW cohesive markers	94.1*	**91.0***	**79.2**	88.6

Table 6: Clustering consensus by voting; statistically significant improvements, compared to using FW only, are marked with '*'

noise diminishes with more data. This result is of clear practical importance, as in real-life situations only a limited amount of data may be available.

The accuracy of supervised classification deteriorates when the size of the underlying logical units (here, chunks) decreases (Kurokawa et al., 2009). We corroborate this observation in the context of clustering, but note that reasonable accuracy (over 70%) can be obtained even with 1000-token chunks (Figure 1b). This further supports the applicability of unsupervised classification of translationese to real-world scenarios.

6 Mixed-domain classification

Poor cross-domain classification results, as described in Section 4, demonstrate that the in-domain discriminative features of translated texts cannot be easily generalized to other, even related, domains. In this section we explore the tension between the discriminative power of domain- and translationese-related properties, in the *unsupervised* scenario. Our underlying hypothesis is that domain-specific features overshadow the features of translationese. The next series of experiments involves (a balanced) combination of various datasets; we excluded the small TED corpus from these experiments to prevent downsampling of other sub-corpora.

6.1 Domain-related vs. translationese-based characteristics

We begin with an investigation of the mutual effect of the domain- and translationese-specific characteristics on the accuracy of clustering. We first merged equal numbers of O and T chunks from two corpora: 800 chunks each from Europarl and Hansard, yielding 1,600 chunks, half of them O and half T. We applied the clustering algorithm of Section 5 to this dataset; the result was a perfect domain-driven separation of all Europarl and Hansard chunks, yielding poor (chance-level) translationese accuracy. In other words, we obtained two clusters, one consisting of Europarl chunks and the other of Hansard chunks, independently of their O-vs.-T status. We repeated the experiment with additional corpus pairs, and further extended it by adding equal numbers of Literature chunks (400 O and 400 T), this time fixing the number of clusters to three. Again, the result was separation by domain: Europarl, Hansard and Literature chunks were grouped into distinct clusters (Table 7, top).

As an additional experiment, we attempted to leave the decision on the "best" number of clusters to the algorithm. To that end, we employed the XMeans clustering procedure (Pelleg and Moore, 2000), which uses KMeans but applies additional statistical cues to decide on the number of clusters that best explain the data. We also applied PCA for dimension reduction prior to XMeans invocation. We repeated both experiments (two- and three-domain mixes) with XMeans, expecting to obtain two and three clusters, respectively. The result is a replication of the more constrained KMeans in three out of four cases (Table 7, bottom).

These observations have a crucial effect on understanding the tension between the domain- and translationese-based characteristics of the underlying texts. Not only are domains accurately separated given a fixed number of clusters, but even when the decision on the number of clusters is left to the clustering procedure, classification into domains explains the data best (as shown by XMeans). Recall that these experiments all rely on the set of function words: topic-independent features, that have been proven effective for telling O from T in both supervised (Section 2) and unsupervised scenarios (Sec-

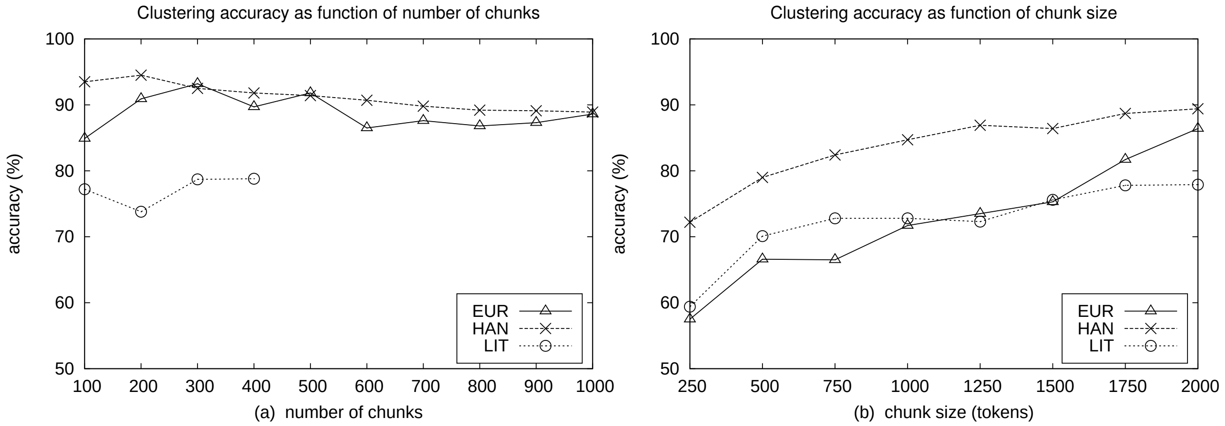

Figure 1: The effect of varying the number of chunks and chunk size (in tokens) on clustering accuracy

method / corpus	EUR + HAN	EUR + LIT	HAN + LIT	EUR + HAN + LIT
KMeans				
accuracy by domain	93.7	99.5	99.8	92.2
accuracy by translation status	50.3	50.0	50.0	–
XMeans				
generated # of clusters	2	2	3	3
accuracy by domain	93.6	99.5	99.9	92.2
accuracy by translation status	50.3	50.0	–	–

Table 7: Clustering a chunk-level mix of Europarl, Hansard and Literature using function words; accuracy by translation status (O vs. T) is reported where applicable (i.e., the outcome constitutes two clusters)

tion 5). The fact that this translationese-oriented feature set yields the results presented in Table 7 clearly demonstrates the dominance of domain-specific properties over the characteristics of translationese.[10]

6.2 Clustering in a mixed-domain setup

Driven by the results of Section 6.1, we turn to explore a methodology for identification of translationese in a mixed-domain setup. We assume that we are given a set of text chunks that come from multiple domains, such that some chunks are O and some are T; the task is to classify the texts to O vs. T, *independently of their domain*. For that purpose, we investigate two approaches: *two-phase* and *flat*. Both methods assume that the number of domains, k, is known (it can be discovered by XMeans, as in Section 6.1, or fixed to a somewhat higher value than estimated in order to capture unsuspected differences within domains). The two-phase method first clusters a mixture of texts into domains (e.g., using KMeans), and then separates each of the resulting (presumably, domain-coherent) clusters into two sub-clusters, presumably O and T. The flat approach applies KMeans, attempting to divide the dataset into $2 \times k$ clusters; that is, we expect classification by domains and by translationese status, simultaneously.

We experimented with two setups: (i) mixture of two datasets out of Europarl, Hansard and Literature (1600 chunks in total); and (ii) mixture of all three of them (2400 chunks in total). We applied both methods to each of the two setups. We invoked PCA prior to clustering in the flat approach; in the two-phase approach, we applied PCA on *raw* data instances that are subject to clustering at each hierarchy level.[11] As our goal is identification of translationese, we define the accuracy of the classification as the ratio of O and T instances classified correctly

[10] Other feature sets yielded similar outcomes.

[11] Note that our two-phase approach differs from the traditional hierarchical clustering in this sense.

method / corpus	EUR + HAN	EUR + LIT	HAN + LIT	EUR + HAN + LIT
Flat	**92.5**	60.7	77.5	66.8
Two-phase	91.3	**79.4**	**85.3**	**67.5**

Table 8: Flat and two-phase clustering of domain-mix using function words

(i.e., we ignore the accuracy of identifying the correct domain).

Table 8 reports the results. Both methods yield similarly high accuracy in the Europarl+Hansard setup, and much lower accuracy in the setup of all three datasets (with a single exception of EUR+LIT). This implies that the difficulty of telling O from T increases as the number of domains in the mixed-domain setup grows. The two-phase approach outperforms the flat one in most cases: the latter attempts to cluster data instances by domain and translation status *simultaneously*, and is therefore potentially more error-prone. As a concrete example, in the Europarl+Literature setup, attempting to produce four clusters, we obtained a single cluster of Europarl chunks and three clusters of Literature chunks. The two-phase approach avoids such pitfalls by explicitly separating the steps of domain- and translationese-based clustering.

Table 8 clearly demonstrates that in a real-world scenario, where a dataset can be assumed to include texts from multiple domains, it is possible to overcome the dominance of domain-related features over translationese-related ones by splitting the task into two. The result is highly accurate identification of translated texts, even in an extremely challenging setup. Compare the results of Table 8 to the *supervised* case (Tables 3, 4): while clustering cannot compete with ten-fold cross-validation results of heterogenous datasets (93–96%), it is far superior to training a classifier on one or more datasets and then using it on a data from a new source (60–64%).

7 Discussion

Distinguishing between original and translated texts has been proven useful for SMT, as awareness to translationese can improve the quality of SMT systems. So far, classifying texts into original vs. translated has been done almost exclusively by supervised methods. In this work we advocate the use of *unsupervised* classification as an effective way to address this task. We demonstrate that simple feature sets, coupled with standard clustering algorithms, a novel cluster labeling technique, and voting among several features, can yield very high accuracy, over 90% in several cases. Using diverse datasets we robustly demonstrate that the approach we advocate is effective for identification of translationese, even when only little data are available, and text chunks are small. We further highlight the dominance of domain-based characteristics of the texts over their translationese-related properties and propose a simple methodology for identification of translationese in a mixed-domain setup. We conclude that the proposed (two-phase) clustering approach is a robust method for distinguishing O from T in heterogenous datasets.

By conducting a series of experiments with unbalanced proportions of O and T texts, we demonstrate that the proposed methodology is also applicable to scenarios where the original and translated data are unevenly distributed.

We applied PCA for dimension reduction and the *tf-idf* weighting scheme with FW throughout all experiments in this work. The latter had a slight positive effect on clustering accuracy in most scenarios, and no impact in some cases. Dimension reduction improved computational efficiency, especially with large feature sets (e.g., character and POS trigrams). However, its effect on clustering accuracy was not uniform: the most prominent improvement (over 15 percent points) was obtained on the TED dataset, while a slight accuracy deterioration was observed in a few cases (e.g., 5 percent points on Europarl with FW). We conclude that while carrying an overall positive value, the application of dimension reduction in similar scenarios calls for further investigation.

8 Conclusion

To the best of our knowledge, this is the first work to extensively explore unsupervised classification of translationese. We only scratched the surface of this research direction. In the future, we intend to explore the robustness of our approach even further, with more datasets in various language pairs. We will first attempt to identify translationese in *French*, using the current dataset (in the reverse direction). We will also experiment with English-German, in both directions, and hopefully also with English-Hebrew, a more challenging setup.

The potential value of unsupervised identification of translationese leaves much room for further exploratory activities. Our future plans include using various datasets and reduced amount of data for LMs compiled for cluster labeling; in particular, we plan to explore the correlation between these two parameters and the scaling factor α used for association of a label with a clustering outcome.

Furthermore, to highlight the contribution of these results to SMT, we plan to replicate the results of Lembersky et al. (2012b, 2013), using *predicted* rather than ground-truth indication of the translationese status of the texts that are used to train SMT systems. We believe that we will be able to show an improvement in the quality of SMT with extremely little supervision.

Acknowledgements

This research was supported by a grant from the Israeli Ministry of Science and Technology. We are grateful to Cyril Goutte, George Foster and Pierre Isabelle for providing us with an annotated version of the Hansard corpus; and to András Farkas[12] and François Yvon for providing us with the Literary corpus. Finally, we are indebted to Noam Ordan, Tamir Hazan, Haggai Roitman and Ekaterina Lapshinova-Koltunski for commenting on earlier versions of this article. All remaining errors and misconceptions are, of course, our own.

References

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5. URL `http://dl.acm.org/citation.cfm?id=1283383.1283494`.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, Forthcoming. doi: http://dx.doi.org/10.1093/llc/fqu047. URL `http://dx.doi.org/10.1093/llc/fqu047`.

Marco Baroni and Silvia Bernardini. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September 2006. URL `http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1`.

Sascha Diwersy, Stefan Evert, and Stella Neumann. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*, pages 174–204. De Gruyter, Berlin, Boston, 2014. URL `http://www.degruyter.com/view/product/207699`.

Dominik Maria Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7): 1858–1860, 2003. URL `http://dblp.uni-trier.de/db/journals/tit/tit49.html#EndresS03`.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund, 1986.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL `http://dx.doi.org/10.1145/1656274.1656278`.

Iustina Ilisei and Diana Inkpen. Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, 2(1-2), 2011.

[12] `http://farkastranslations.com`

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL `http://dx.doi.org/10.1007/978-3-642-12116-6`.

Zahurul Islam and Alexander Mehler. Customization of the Europarl corpus for translation studies. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), may 2012. ISBN 978-2-9517408-7-7.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 2nd edition, 2002.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86. AAMT, 2005. URL `http://mt-archive.info/MTS-2005-Koehn.pdf`.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for Europe. In *Proceedings of the Twelfth Machine Translation Summit*, pages 65–72, 2009.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1132`.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pages 81–88, 2009.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1034`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France, April 2012a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E12-1026`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825, December 2012b. URL `http://dx.doi.org/10.1162/COLI_a_00111`.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4):999–1023, December 2013. URL `http://dx.doi.org/10.1162/COLI_a_00159`.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, January 1991. ISSN 0018-9448. doi: 10.1109/18.61115. URL `http://dx.doi.org/10.1109/18.61115`.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, March 1982. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489. URL `http://dx.doi.org/10.1109/TIT.1982.1056489`.

Gerard Lynch and Carl Vogel. Towards the automatic detection of the source language of a literary translation. In *Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics: Posters*, pages 775–784, 2012. URL `http://aclweb.org/anthology/C12-2076`.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association* *for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2001. URL http://papers.nips.cc/paper/2092-on-spectral-clustering-analysis-and-an-algorithm.pdf.

Sergiu Nisioi and Liviu P. Dinu. A clustering approach for translationese identification. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing, RANLP 2013*, pages 532–538. RANLP 2011 Organising Committee / ACL, September 2013.

Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2. URL http://dl.acm.org/citation.cfm?id=645529.657808.

Marius Popescu. Studying translationese at the character level. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *Proceedings of RANLP-2011*, pages 634–639, 2011.

Ella Rabinovich and Shuly Wintner. The Haifa corpus of translationese. Unpublished manuscript, Forthcoming.

Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *KDD-2000 Workshop on Text Mining*, August 2000.

Hans van Halteren. Source language markers in EUROPARL translations. In Donia Scott and Hans Uszkoreit, editors, *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 937–944, 2008. ISBN 978-1-905593-44-6. URL http://www.aclweb.org/anthology/C08-1118.

Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, April 2015.